Clinical Chemistry 55:4
698–708 (2009)

Management of Gene Promoter Mutations in Molecular Diagnostics

Karen M. K. de Vooght,[1*] Richard van Wijk,[1] and Wouter W. van Solinge[1]

BACKGROUND: Although promoter mutations are known to cause functionally important consequences for gene expression, promoter analysis is not a regular part of DNA diagnostics.

CONTENT: This review covers different important aspects of promoter mutation analysis and includes a proposed model procedure for studying promoter mutations. Characterization of a promoter sequence variation includes a comprehensive study of the literature and databases of human mutations and transcription factors. Phylogenetic footprinting is also used to evaluate the putative importance of the promoter region of interest. This in silico analysis is, in general, followed by in vitro functional assays, of which transient and stable transfection assays are considered the gold-standard methods. Electrophoretic mobility shift and supershift assays are used to identify trans-acting proteins that putatively interact with the promoter region of interest. Finally, chromatin immunoprecipitation assays are essential to confirm in vivo binding of these proteins to the promoter.

SUMMARY: Although promoter mutation analysis is complex, often laborious, and difficult to perform, it is an essential part of the diagnosis of disease-causing promoter mutations and improves our understanding of the role of transcriptional regulation in human disease. We recommend that routine laboratories and research groups specialized in gene promoter research cooperate to expand general knowledge and diagnosis of gene-promoter defects.

© 2009 American Association for Clinical Chemistry

Gene expression is regulated at many levels, including chromatin packing, histone modification, transcription initiation, RNA polyadenylation, pre-mRNA splicing, mRNA stability, and translation initiation. An important part of regulation, however, is believed to occur at the level of transcription initiation *(1)*. During the past few years, much progress has been made in understanding the basis of transcriptional regulation. Transcription factors (TFs),[2] chromatin-modifying enzymes, and TFs unite to activate genes and are recruited in a precise order to promoters. The timing of the activation of transcription and the ordered recruitment of factors to promoters are the engines that, at the right moment and for the right duration of time, drive transcriptional regulation of each gene throughout the cell's life-span *(2)*. Failure in timing or recruitment of TFs may affect transcriptional regulation of a gene, putatively leading to disease.

In this review we focus on sequence variations in the promoter region as a putative cause of disturbed transcriptional regulation leading to disease. Not every promoter sequence variation affects transcriptional regulation. Depending on the location and the nature of the genetic defect, a mutation in the promoter region of a gene may disrupt the normal processes of gene activation by disturbing the ordered recruitment of TFs at the promoter. As a result a promoter mutation can decrease or increase the level of mRNA and thus protein.

The effect of promoter mutations can be very subtle. In addition, promoter mutation analysis is complex, and the assays that are needed to investigate the functional relationship between the mutation and disease are laborious and difficult to perform. Therefore, thorough studies of promoter mutations are scarce and often confined to research laboratories.

The Promoter of a Gene

The promoter (Fig. 1), a regulatory region of DNA located upstream of a gene, plays an important role in transcriptional regulation. The core promoter, a loosely defined region (approximately between nucleotides −40 and +50 from the transcriptional start site

[1] Department of Clinical Chemistry and Haematology, Laboratory for Red Blood Cell Research, University Medical Center Utrecht, Utrecht, the Netherlands.
[*] Address correspondence to this author at: Department of Clinical Chemistry and Haematology, Laboratory for Red Blood Cell Research, University Medical Center Utrecht, Postbus 85500, 3508 GA, Utrecht, the Netherlands. Fax +31 88 7555418; e-mail k.devooght@umcutrecht.nl.

Received November 19, 2008; accepted January 28, 2009.
Previously published online at DOI: 10.1373/clinchem.2008.120931

[2] Nonstandard abbreviations: TF, transcription factor; TSS, transcriptional start site; TFBS, transcription factor binding site; HGMD, human gene mutation database; SNP, single nucleotide polymorphism; EMSA, electrophoretic mobility shift assay.

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

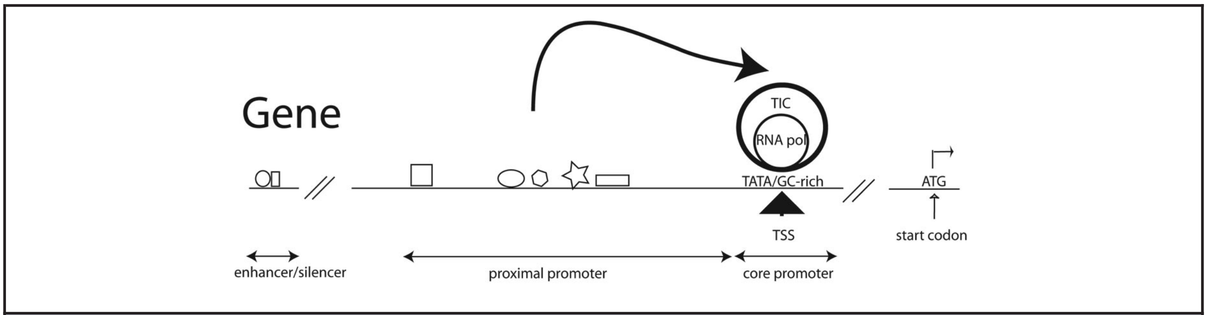

Fig. 1. Schematic overview of the different elements of a general promoter.
The core promoter directs low-level transcription and contains binding sites for general transcription factors and RNA polymerase II. The proximal promoter contains multiple binding sites for transcription factors, which cooperatively stimulate transcriptional activity. Transcription factors are indicated by different geometric shapes. TIC, transcription initiation complex; RNA pol, RNA polymerase.

[TSS]), directs low-level transcription. The core promoter region contains binding sites for general TFs and RNA polymerase II. These general TFs, such as TFIID, TFIIA and TFIIB, assemble on the core promoter in an ordered fashion to form a transcription–initiation complex, which directs RNA polymerase II to the TSS *(3)*. The core promoter may also contain other elements such as the TATA box, which is the binding site for a subunit of TFIID. This TATA box has the consensus-binding sequence 5′-TATAAA-3′ and is characteristic for tissue-specific genes, the expression of which is restricted to a limited number of cells *(4)*. Housekeeping genes, the expression of which is ubiquitous, usually lack TATA boxes and instead contain GC-rich sequences.

The assembly of general TFs on the core promoter is sufficient to direct low levels of transcription, a process generally referred to as basal transcription. Transcriptional activity is greatly stimulated by a second class of TFs, termed activators. In general, activators are sequence-specific DNA-binding proteins whose recognition sites are present in the proximal promoter. The proximal promoter is the region immediately upstream, up to a few hundred base pairs, from the core promoter, and typically contains multiple binding sites for TFs *(1)*.

In contrast to the core and proximal promoter, enhancers are regulatory DNA sequences that may be located 5′ or 3′ to or within an exon or intron of a gene. Enhancer function is by definition independent of position and orientation. Enhancers are considered to act via a DNA-loop, whereby the enhancer and core promoter are brought into close proximity by "looping out" the intervening DNA *(1)*. Whereas there are common motifs in core and proximal promoters, enhancers do not contain many distinctive sequence motifs. Therefore they cannot easily be identified on the basis of their DNA sequence alone.

Sequence-specific elements that confer a negative (i.e., silencing or repressing) effect on the transcription of a target gene are called silencers. They generally have the same features as enhancers. In addition, the locus control region is a group of regulatory elements involved in regulating an entire locus or gene cluster. Locus control regions direct tissue-specific, physiological expression of a linked gene in a manner that is position independent and copy-number dependent and are composed of multiple cis-acting elements, including enhancers and silencers *(1, 3)*.

Many classes of TFs, which can be distinguished from each other by different DNA-binding domains, have been described. Examples of activator families include those containing a cysteine-rich zinc finger, homeobox, helix-loop-helix, basic leucine zipper, forkhead, or ETS DNA-binding domain *(3)*. The TF-binding sites (TFBS) are generally small, in the range of 6–12 bp, although binding specificity is usually dictated by no more than 4–6 positions within the site *(1)*. The TFBS for a specific activator is therefore typically described by a consensus sequence in which certain positions are relatively constrained whereas others are more variable.

In September 2003, the National Human Genome Research Institute launched the ENCODE (encyclopedia of DNA elements) project to identify all functional elements in the human genome by using a mix of different experimental and computational approaches. During its pilot phase, the project focused on approximately 1% of the human genome sequence (ENCODE Project Consortium 2004: http:/www.genome.gov/10005107). One of the most surprising findings was that more than half of the genes use a tissue-specific

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

Table 1. Frequently used Web-based resources for in silico promoter analyses *(32)*.[a]

Resource name	URL	Information outcome
Genome browsers		
NCBI Ensembl	http://www.ensembl.org	Gene sequence, polymorphic variations, exon information, phylogenetic footprinting
OMIM (Online Mendelian Inheritance in Man)	http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM	Gene information, promoter information, mutations
National Center for Biotechnology Information (NCBI) Entrez Nucleotide	http://www.ncbi.nlm.nih.gov/sites/entrez?db=Nucleotide	Gene sequence, promoter sequence
Mutation database		
HGMD (Human Gene Mutation Database)	http://www.hgmd.cf.ac.uk/ac/index.php	Mutations
Promoter predictions		
PromoterInspector	http://www.genomatix.de/online_help/help_gems/PromoterInspector_help.html	Promoter prediction
FirstEF	http://rulai.cshl.edu/tools/FirstEF	Promoter prediction
DBTSS (Database of Transcriptional Start Sites)	http://dbtss.hgc.jp/index.html	Transcriptional start site
TF-binding profile database		
TRANSFAC®	http://www.gene-regulation.com/pub/databases.html#transfac	TF-binding site (matrix), TFs
TF-binding site prediction		
TESS (Transcription Element Search System)	http://www.cbil.upenn.edu/cgi-bin/tess/tess	TF-binding site prediction
Match™	http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi	TF-binding site prediction

[a] Please note that this list is not intended to be comprehensive. For a more extensive list, see Wasserman et al. *(32)*.

and often unannotated set of exons outside the current boundaries of the annotated genes *(5)*. In some genes the promoters of other neighboring genes are used in specific cells and/or developmental stages. In addition, 5′ untranslated regions have been shown to contain critical regulatory elements *(6, 7)*. These findings contribute to the opinion that transcriptional regulation is complex and therefore difficult to study.

Location of the Promoter

To understand the mechanisms of transcriptional regulation of a certain gene, knowledge of the exact location of the promoter(s) and possible enhancers and silencers is necessary to relate promoter mutations to disease. In addition, this information is required to design correct promoter reporter vectors for transfection assays, the gold standard assay for investigating functional importance of promoter mutations (see below). Identifying the promoter of a specific gene poses a challenge, because core promoters are often located far upstream of the first coding exon. Furthermore, at least half of the mammalian genes are regulated by more than one promoter to enable tissue-specific regulation *(8)*. Fortunately, the promoters of many genes have recently been identified, and some of the most important TFBSs have been characterized *(1)*. Promoter prediction programs (e.g., PromoterInspector, FirstEF) may be used to identify and locate the promoter if this information is not available (Table 1). These programs are frequently modified to make them more accurate and efficient.

It may be challenging to determine which region of the promoter should be screened for regulatory mutations. Recently, 5′ untranslated regions have been shown to contain several critical regulatory elements *(6)*. Therefore it seems appropriate to start screening immediately upstream of the translation initiation site (which starts with the nucleotide sequence ATG). Evaluation of entries in the Human Gene Mutation Database (HGMD) *(9)* reveals that most registered regulatory mutations are located between nucleotides +50 and −500 from the TSS of a gene. Rockman et al. *(10)* analyzed the distribution of functional single-

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

Table 2. Examples of mutations in transcriptional regulatory elements associated with human diseases.

Disease	Affected gene	Mutation (disrupted regulatory element)	Reference
β-thalassemia	*HBB*	Numerous (TATA-box, CACCC box, EKLF)	Hardison et al. *(15)*
Bernard-Soulier syndrome	*GP1BB*	–133A→G (GATA-1)	Ludlow et al. *(33)*
Pyruvate kinase deficiency	*PKLR*	–72A→G (GATA-1) −83G→C (PKR-RE1)	Manco et al. *(34)*, Van Wijk et al. *(35)*
Familial hypercholesterolemia	*LDLR*	Numerous (Sp-1, SRE repeat, FP1, FP2)	www.ucl.ac.uk/ldlr
Hemophilia B	*F9*	−20T→A (HNF-4), −6G→A and −6G→C (C/EBP)	Crossley and Brownlee *(36)*, Reijnen et al. *(37)*

nucleotide polymorphisms (SNPs) (see also below) in the human promoter region and showed that the first 500 nucleotides upstream of the TSS indeed contained most of the functional SNPs (59%). However, a substantial fraction was found further upstream; 13% were more than 1 kb upstream, and another 13% were located 3′ to the TSS. The authors even reported that 2 SNPs (1.4%) occurred even more than 10 kb upstream of their TSS. There is, therefore, a spatial distribution with respect to sequence variations affecting transcriptional regulation *(10)*, although there is a bias toward the immediate 5′ flanking sequence. These findings indicate that in case of a suspected regulatory mutation causing disease without alterations in the proximal or core promoter region, the upstream region is likely to be a good target for further analysis. Confirming this assumption are reports that mutations in upstream promoter regions, such as enhancers, silencers, and locus control regions, are associated with disease *(11)*.

Significance of Promoter Mutations in Human Disease

PROMOTER MUTATIONS IN DISEASE

Some 1% of single base-pair substitutions causing human genetic disease occur within gene promoter regions, where they disrupt the normal processes of gene activation and transcriptional initiation and usually decrease or increase the amount of mRNA and thus protein *(12)*. Promoter mutations can alter or abolish the binding capacity of cis-acting DNA-sequence motifs for the trans-acting protein factors that normally interact with them *(12)*. Examples of promoter mutations causing disease include β-thalassemia, Bernard-Soulier syndrome, pyruvate kinase deficiency, familial hypercholesterolemia, and hemophilia (Table 2) *(1)*. The contribution of promoter mutations to the total of disease-causing mutations is unclear, however. For instance, the majority of missense mutations cause a qualitative defect that is fairly easy to identify. Sometimes, mutant alleles even act as dominant alleles, because the affected protein may antagonize remaining normal protein. In contrast, promoter mutations may cause small quantitative defects, which may be hard to detect. Even if the promoter of an autosomal gene is completely downregulated as result of mutation, half of the normal amount of protein is present, which is often enough to prevent severe disease.

Because there are few reports about the incidence of promoter mutations, we studied the HGMD *(9, 13)*. To date, this database contains a total of 73 411 registered mutations (assembly date September 2007), of which 1.5% are regulatory. An example of a thoroughly studied gene, which has been a model gene for studying mechanisms of transcriptional regulation, is the hemoglobin, beta (*HBB*)[3] gene. The HGMD database contains a total of 490 entries for *HBB*, of which 234 (48%) are missense/nonsense mutations, 28 (6%) promoter mutations, and 9 (2%) other (3′) regulatory mutations. The first regulatory mutation entry was that of a single base change (–28A→C) in the TATA box of the *HBB* gene, which caused β-thalassemia in a Kurdish Jewish individual in 1982 *(14)*. This modification of the TATA box was the first ever found in association with a genetic disorder. Approximately 10 of 28 registered *HBB* promoter mutations have been studied by use of functional transfection assays *(15)*. An example of a gene in which regulatory mutations have only recently been identified is the cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) (*CFTR*) gene. This gene was identified in 1989, and the catalogue of mutations now exceeds 1564 in number (www.genet.sickkids.on.ca/cftr) but contains only 8 promoter mutations (0.52%). The first DNA defect, the well-known ΔF508 deletion

[3] Genes: *HBB*, hemoglobin, beta; *CFTR*, cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7); *GP1BB*, glycoprotein 1b (platelet), beta polypeptide; *PKLR*, pyruvate kinase, liver and RBC; *LDLR*, low density lipoprotein receptor; *F9*, coagulation factor IX; *PPOX*, protoporphyrinogen oxidase; *luc*+, a gene of the firefly.

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

causing cystic fibrosis, was reported in 1989 *(16)*, whereas Bienvenu et al. *(17)* reported the first regulatory mutation (–741T→G) almost 6 years later. In contrast to *HBB*, only 1 of the 8 catalogued *CFTR* promoter mutations has been characterized by use of functional transfection assays *(18)*. Although the relevance of promoter mutations in cystic fibrosis is unknown, these observations suggest that the number of putative *CFTR* promoter mutations is underestimated. As in *CFTR*, promoter mutations may have been overlooked in other genes. As a result, it is difficult to assess the general incidence of disease-causing promoter mutations.

POLYMORPHIC PROMOTER SEQUENCE VARIATIONS IN DISEASE

In general, polymorphic sequence variations are considered to be rather harmless, especially if located in noncoding parts of a gene. The role of polymorphisms in determining susceptibility to disease traits is the subject of much research effort, but it often remains unclear whether the polymorphisms are themselves functionally relevant or just linked to another causative mutation *(19)*. The term polymorphism has been defined as a "Mendelian trait that exists in the population in at least two phenotypes, neither of which occurs at a frequency of <1%" *(12)*. Polymorphisms are not rare, being distributed thorough the human genome at a frequency of 1 in 200 to 1 in 1000 bp *(20)*. Polymorphisms that occur in the promoter may affect gene expression and may thus have the potential to be of phenotypic or even of pathological significance *(12)*. An increasing number of promoter polymorphisms have been characterized by functional studies. Some may well be pathologically important, e.g., those in the genes coding for plasminogen activator inhibitor type 1, tumor necrosis factor α, apolipoprotein AI, lipoprotein lipase, and interleukin 6 *(12)*.

Current epidemiological investigations, in which large amounts of SNPs are studied in relation to disease, are revealing considerable numbers of putative functional promoter SNPs. However, a causal link between these promoter polymorphisms and disease is often absent, because these studies generally lack functional promoter assays. Without functional promoter assays it is incorrect to state that a certain promoter sequence variation causes disease in vivo; another regulatory mutation linked to the identified polymorphism may be the one affecting promoter activity, thereby causing disease.

Techniques for Promoter Analysis

IN SILICO ANALYSIS OF PROMOTER MUTATIONS

Literature and databases such as HGMD, National Center for Biotechnology Information (NCBI) Ensembl, and the online version of Mendelian Inheritance of Man (OMIM) can be used as a first step to investigate if an identified promoter sequence variation is known, associated with disease, and previously functionally characterized (Tables 1 and 3). DNA polymorphisms are often not catalogued in these databases unless they exhibit sufficiently strong phenotypic association *(21)*.

The next step is to use in silico analysis to investigate whether the sequence variation is disrupting or creating a putative TFBS (Table 3). Experimental data regarding the specific binding sites of most well-characterized TFs have been compiled in databases such as TRANSFAC (Table 1) *(22)*. In these databases experimentally determined TFBSs are used to calculate a probability score for nucleotides on a specific position in a consensus TFBS (site matrix). Programs such as TESS (transcription element search software) (Table 1) *(23)* are able to compare a genomic sequence input to all matrices in TRANSFAC and report a list of potential TFBSs based on a statistical match between a region in the sequence and a site matrix. This analysis is, however, often encumbered by the prediction of a large number of putative TFBSs, a significant fraction of which will not be involved in transcriptional regulation of the gene. This situation may be attributable to the quality of the data used to build the TFBS matrices *(24)* and discrepancies that occur owing to in vivo absence or inactivity of a TF or cofactor, or to condensed local chromatin *(25)*. In addition to these false-positive problems, the comprehensiveness of the databases is also an issue; not all DNA-binding TFs have been identified, and even for some known factors, binding specificity has not yet been fully characterized *(1)*.

Phylogenetic footprinting (Table 3), the comparison of the sequence of interest with the homologous region in other species, is used to investigate the putative functional relevance of a promoter sequence variation. The rationale behind this process is that nucleotides within binding sites are more likely to be conserved by natural selection. Although there is abundant evidence that conserved regions do, indeed, often contain functional regulatory motifs; this correlation does not always exist because not all TFBSs are conserved among species. Finally, some of the most important transcriptional regulatory elements relevant to normal human development and disease may not be highly conserved. Instead they may be restricted to only humans or primate relatives *(1)*.

FUNCTIONAL PROMOTER ASSAYS

A promoter mutation that putatively causes disease must be characterized to assess the relevance of the DNA sequence variation in relation to the disease. Proper analysis demands proof that the mutation sig-

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

Table 3. Properties of techniques involved in promoter analysis.

Method	Aim	Description	Feasible in general laboratory	Main advantages	Main disadvantages	References
In silico analysis						
Literature and database search	To investigate prevalence, association with disease and previous functional characterization	Literature and databases are used to investigate relevance of promoter mutation	Yes	Fast, easy, low cost	Publication bias, absence of information in case of unknown mutations, polymorphisms often not catalogued	Antonarakis et al. *(21)*
TF database search	To investigate if the sequence variation is disrupting or creating a putative TFBS	Programs compare genomic sequence input to known binding sequences and report list of potential TFBSs	Yes	Fast, easy, low-cost assessment of putative relevance of mutation possible	In silico vs in vivo discrepancies, not all DNA-binding TFs have been identified or fully characterized	Maston et al. *(1)*, Wray et al. *(25)*
Phylogenetic footprinting	To investigate putative relevance of promoter sequence variation	Comparison of sequence with homologous region in other species.	Yes	Fast, easy, low cost, assessment of putative relevance of mutation possible	Correlation between conservation and functional activity does not always exist, not all TFBSs are conserved among species	Maston et al. *(1)*
Functional promoter assays						
Transient transfection assays	To investigate if promoter mutation alters promoter activity in vitro	Promoter constructs are cloned into plasmid upstream of reporter gene and transiently transfected into cultured cells; reporter activity correlates with promoter activity	No, cloning and cell culturing facilities have to be available	Prediction of in vivo promoter activity, relatively easy to perform and less time-consuming compared to other functional promoter assays	Difficulties capturing all regulatory elements in single-reporter construct, in vitro–in vivo discrepancies due to differences in chromatin context and due to tissue- or developmental-stage–specific expression, plasmid DNA exists in artificial configuration leading to inactivity of regulatory elements	Maston et al. *(1)*, Knight *(19)*
Stable transfection assays	To investigate if promoter mutation alters promoter activity in vitro	Promoter constructs are cloned into plasmid upstream reporter gene and stably transfected into cultured cells; reporter activity correlates with promoter activity.	No, cloning and cell culturing facilities have to be available	Prediction of in vivo promoter activity, selection of transfected cells based on drug-resistance gene, natural chromatin environment and copy number, more sensitive ad robust than transient transfection assays.	Difficulties capturing all regulatory elements in single reporter construct, in vitro–in vivo discrepancies due to differences in chromatin context and due to tissue- or developmental-stage–specific expression, more technically demanding than transient transfection assays	Maston et al. *(1)*, Knight *(19)*
Transgenic expression assays	To investigate if promoter mutation alters promoter activity in vivo	Promoter constructs are cloned into reporter assay system and injected in fertilized oocytes;. in vivo gene expression correlates with reporter-gene expression in transgenic organism	No, cloning and cell-culturing facilities have to be available	Better prediction of in vivo promoter activity than with other functional assays, monitoring reporter-gene expression through entire development of the organism	Laborious and time-consuming	Nobrega et al. *(27)*
DNA-TF–binding assays						
EMSA	To investigate protein binding to promoter in vitro	Proteins recognizing a given promoter sequence are identified by incubating labeled DNA probe with nuclear extract; upon electrophoresis free labeled–probe molecules separate from protein-bound molecules	Yes, when using commercially available nuclear extracts and nonradiolabeled probes	Clear-cut method, sensitive, detection and characterization of specific protein–DNA interactions, identification of TFs by supershift assays, commercial assays available, location of TFBS determined by mutagenesis, detection of protein–DNA interactions on a small fragment (20–30 bp)	Protein–DNA interaction should maintain during gel electrophoresis, length of probe is critical, titration of nonspecific competitor DNA difficult, DNA region of interest must be known, only short probes can be tested	Knight *(19)*, Carey and Smale *(28)*
DNase I–footprinting assay	To investigate protein binding to promoter in vitro	Partial digestion of ^{32}P-labeled fragment by DNase I is followed by denaturing acrylamide gel electrophoresis; nucleotides bound by protein are protected from cleavage, producing "footprint" in ladder of labeled DNA fragments, allowing specific localization of the site of protein–DNA interaction	No, ^{32}P facilities must be available	Useful for scanning large DNA fragments (50–200 bp) for protein–DNA interactions	Only broad indication of binding site, DNA backbone affects cleavage efficiency, effective primarily when protein is at high concentrations, $MgCl_2$ and $CaCl_2$ (for DNase I activity) may disrupt specific interactions, less useful for investigation of specific promoter mutation	Knight *(19)*, Carey and Smale *(28)*, Galas and Schmitz *(38)*
Chromatin-immunoprecipitation (ChIP) assay	To investigate protein binding to promoter in vivo	DNA-binding proteins are crosslinked to target sites in growing cells; cells are lysed, DNA cleaved; protein–DNA complexes are purified by immunoprecipitation with antibodies directed against the DNA-binding protein; immunoprecipitate is analyzed for presence of regulatory element	Yes, when using commercially available nuclear extracts	"Visualizing" of in vivo interaction between a specific protein and regulatory element, commercial ChIP assay kits (shortening optimization procedures) available	Technically challenging, TF must be known, high-quality antibodies needed, optimizing shearing conditions difficult, careful titration of DNase I necessary	Carey and Smale *(28)*

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

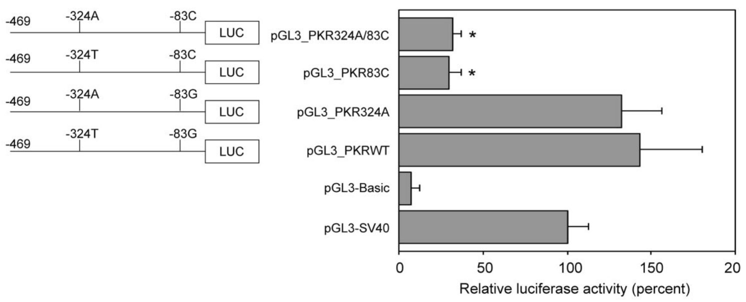

Fig. 2. Results obtained by use of transient transfection assays.

The −324T→A mutation in the erythroid-specific promoter of the pyruvate kinase, liver and RBC (*PKLR*) gene does not affect promoter activity compared with the wild-type construct. In contrast, the −83G→C strongly reduces in vitro promoter activity. We concluded that the −83G→C mutation in the erythroid-specific promoter of *PKLR* strongly downregulates promoter activity in vitro *(35)*. Reprinted with permission from Van Wijk et al. *(35)*. *Statistically significant ($P < .05$). LUC, the pGL3-Basic vector (Promega Corporation) with *luc*+ gene (a gene of the firefly).

nificantly alters promoter activity in vitro (functional assays, Table 3). One of the more versatile functional tests is based on the use of reporter gene assays. In these assays, the region of DNA to be tested for regulatory activity is cloned into a plasmid upstream of an easily assessable reporter gene, such as the genes coding for chloramphenicol acetyltransferase, β-galactosidase, green fluorescent protein, or luciferase *(26)*. The resulting wild-type and mutant constructs are then transfected (either transiently or stably) into cultured cells, and the activity of the reporter gene is measured to determine if the promoter mutation alters reporter gene expression (see for an example Fig. 2). Cotransfection of a control reporter plasmid is used to correct for transfection efficiency within or between transfection experiments. More sophisticated testing of upstream regulatory elements is performed by constructing transgenic organisms and monitoring reporter gene expression through the entire development of the organism *(27)*. Compared to constructing transgenic organisms, transient transfection assays are much easier to perform, less time-consuming, and more feasible in laboratories with only limited cell-culturing facilities. One of the restrictions is that regulatory elements can be widely dispersed and difficult to capture in a single reporter construct. Another concern is that the plasmid DNA is placed in an artificial environment, which may lead to inactivity or dysregulation of regulatory elements *(19)*. A third drawback is that the in vivo activity of a reporter gene may fail to reproduce the expression pattern of its endogenous equivalent owing to differences in chromatin context. Finally, a given upstream regulatory element may, in practice, be used only for restricted purposes, such as those specific for certain tissues or developmental stages. If the cell culture system used to assay the reporter gene activity does not match the physiological conditions under which the regulatory element is normally active, differences in promoter activity between wild-type and mutant constructs may not be detected *(1)*. Despite the limitations, reporter gene assays remain the most accurate means available to investigate the functional consequences of a promoter mutation.

DNA-TF–BINDING ASSAYS

In addition to functional promoter assays, it is essential to demonstrate that the interaction of a putative TF with the DNA sequence of interest is affected by the promoter mutation (DNA-TF–binding assays, Table 3) *(28)*. Several methods have been used for in vitro detection and characterization of protein–DNA interactions, including electrophoretic mobility shift assay (EMSA) and DNase I–footprinting assays (Table 3). EMSA is by far the most commonly used assay, mainly because it provides a relatively simple, rapid, and extremely sensitive technique for the detection and characterization of specific protein–DNA interactions *(29)*. EMSA is based on the principle that a protein–DNA complex migrates more slowly through a native gel than the corresponding free DNA. Proteins within a nuclear extract that specifically recognize a given promoter sequence can be identified by incubating a small radiolabeled DNA probe with the extract to allow the formation of protein–DNA complexes. Application of the mixture to a native polyacrylamide gel and subsequent electrophoresis, will separate the free radiolabeled probe molecules from the protein-bound molecules *(28)*. Free DNA and DNA–protein complexes are then detected by autoradiography or phosphorimager analysis (Fig. 3). Differences in binding pattern be-

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

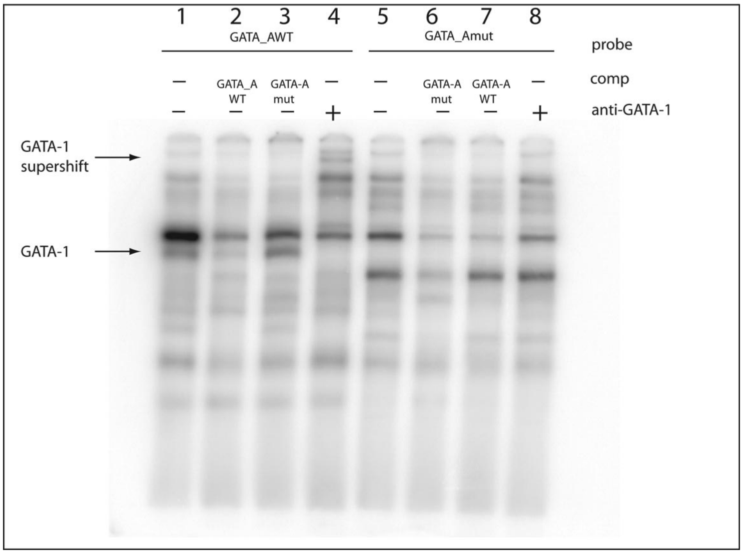

Fig. 3. Results obtained by use of EMSA and supershift assay.

Incubation of a wild-type (WT) protoporphyrinogen oxidase (*PPOX*) probe with K562 nuclear extract results in a specific DNA–protein complex. This complex is absent in case of a GATA-1 mutated (mut) *PPOX* competitor (comp) probe. The DNA–protein complex supershifts after addition of anti–GATA-1. We concluded that GATA-1 binds 1 of the GATA-binding motifs in exon 1 of *PPOX* in vitro *(6)*. Reprinted with permission from de Vooght et al. *(6)*.

tween the wild-type and mutant radiolabeled probes are indicative of TFs interacting with the DNA sequence of interest. Competition studies with nonlabeled wild-type and mutant competitor probes are used to test the specificity of DNA–protein interactions. Commercial kits for performing EMSA without the need for radiolabeled probes have recently become available. The putative TF candidate can be further identified by use of an antibody directed against this protein. After electrophoresis, binding of this antibody to the DNA–protein complex results in a more slowly migrating or completely disappearing DNA–protein complex (supershift assay, Fig. 3). In case of an unknown protein interacting with the DNA sequence of interest, protein purification experiments must be performed first.

One advantage of EMSA is that it is analytically sensitive and can reveal a specific protein–DNA complex even when the protein is present at low concentrations. A disadvantage is that the protein–DNA interaction has to be maintained during gel electrophoresis. Some protein-DNA complexes are not sufficiently strong to last during the typical 2- to 4-h electrophoresis time period. In addition, probes should be long enough to support the forming of stable protein-DNA interactions, and relatively large concentrations of nonspecific competitor DNA, such as poly(dI:dC), are often needed to increase specificity *(28)*.

Chromatin immunoprecipitation assays demonstrate the in vivo relevance of TF binding. In brief, in these assays growing cells are treated with formaldehyde to crosslink DNA-binding proteins to their target sites. Cells are then lysed, and the DNA is cleaved into fragments by digestion with a restriction enzyme or by ultrasonic shearing. Protein–DNA complexes are purified by immunoprecipitation with antibodies directed against the DNA-binding protein of interest. To determine whether the protein was crosslinked to the putative TFBS, antibody-binding is neutralized, proteins are digested by proteinase-K treatment, and DNA is analyzed by PCR for the presence of a DNA fragment encompassing the regulatory element *(30)*. The principal strength of the in vivo crosslinking assay is that it is the only method currently available for directly "visualizing" an in vivo interaction between a specific protein and a regulatory element *(28)*. A limitation of the approach is that it is technically challenging and that the putatively interacting TF must be known. The method requires high-quality antibodies capable of recognizing the fixed, target-bound TF, and optimization of chromatin-shearing conditions can be difficult. Fortunately, commercial chromatin immunoprecipitation assay kits have recently become available. These kits shorten optimization procedures, making these assays accessible for less experienced laboratories.

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

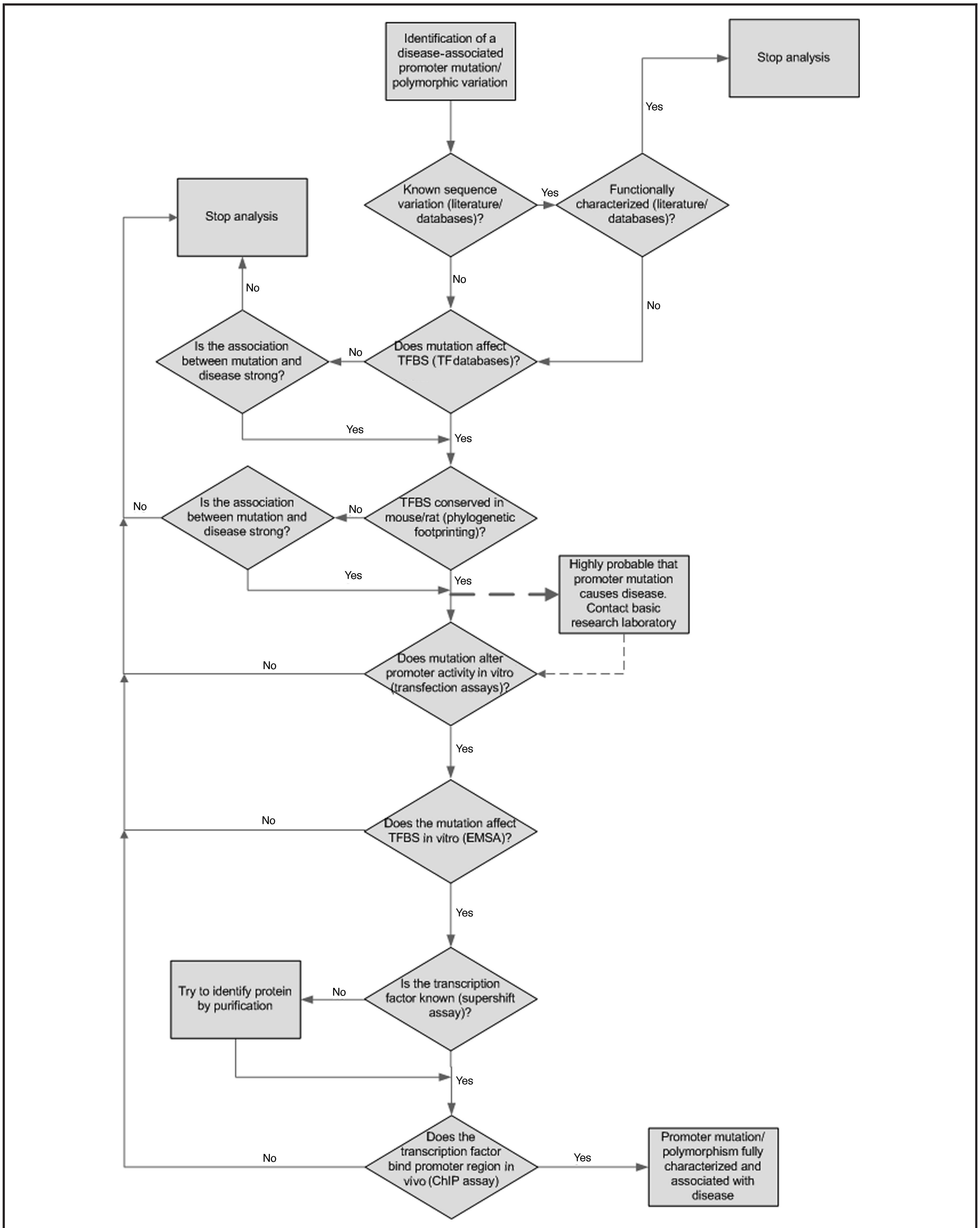
Identification of a disease-associated promoter mutation/ polymorphic variation
Known sequence variation (literature/ databases)?
Yes
Functionally characterized (literature/ databases)?
Yes
Stop analysis
No
No
Does mutation affect TFBS (TF databases)?
No
Is the association between mutation and disease strong?
No
Stop analysis
Yes
Yes
TFBS conserved in mouse/rat (phylogenetic footprinting)?
No
Is the association between mutation and disease strong?
No
Yes
Yes
Highly probable that promoter mutation causes disease. Contact basic research laboratory
Does mutation alter promoter activity in vitro (transfection assays)?
No
Yes
Does the mutation affect TFBS in vitro (EMSA)?
No
Yes
Is the transcription factor known (supershift assay)?
No
Try to identify protein by purification
Yes
Does the transcription factor bind promoter region in vivo (ChIP assay)
No
Yes
Promoter mutation/ polymorphism fully characterized and associated with disease

Fig. 4. Flow chart of the proposed standardized procedure for investigating promoter mutations and polymorphisms.

Different aspects of promoter analysis are put into a decision tree for characterization of promoter mutations and polymorphic-promoter sequence variations. ChIP, chromatin immunoprecipitation.

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

PROCEDURE FOR ANALYZING PROMOTER MUTATIONS AND POLYMORPHISMS

In general, the characterization of a detected promoter variation can be performed according to the proposed standardized procedure displayed in Fig. 4. This flowchart is compatible with most published promoter mutation studies. In silico analysis is relatively quick and easy and can also be performed by less experienced laboratories. Functional promoter assays and TF-binding assays are more difficult and laborious to perform.

Concluding Remarks

Several studies have identified disease-causing cis-regulatory mutations *(31)*. The main reasons why promoter analysis is not performed on a regular basis are: (*a*) promoter mutations can be properly analyzed only by laborious functional or biochemical tests, (*b*) the location of the promoter is frequently not well defined, (*c*) the significance of promoter mutations is difficult to interpret, and (*d*) the effect of promoter mutations is considered to be too mild to cause disease. As a result, interpretation of promoter mutations is difficult and often not a feasible way to gain strong conclusive results with regard to the clinical effect of the identified mutation.

Analysis of promoter mutations is important because it improves the diagnosis of disease-causing promoter mutations and also expands our understanding of the role of transcriptional regulation in human disease. To enhance the diagnosis of disease caused by mutations in the promoter region of a gene and to speed up procedures in basic promoter-research laboratories we need better prediction programs and dedicated easy-to-perform functional promoter assays as well as TF-binding assays for analyzing promoter mutations. Pending these advances, we believe that clinical laboratories should team up with research groups specializing in gene-promoter research. This is a 2-way street: routine laboratories can translate results obtained by research laboratories into diagnostic tools, whereas research groups specializing in gene-promoter research depend on the identification of regulatory mutations in patients to improve knowledge of transcriptional regulation of the gene of interest and the role of transcriptional regulation in disease.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References

1. Maston GA, Evans SK, Green MR. Transcriptional regulatory elements in the human genome. Annu Rev Genomics Hum Genet 2006;23:23.
2. Cosma MP. Ordered recruitment: gene-specific mechanism of transcription activation. Mol Cell 2002;10:227–36.
3. Latchman DS. Eukaryotic transcription factors. 3rd ed. London: Academic Press; 1998. 360 p.
4. Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. Ann Rev Biochem 1981;50: 349–83.
5. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, et al. Prominent use of distal 5′ transcription start sites and discovery of a large number of additional exons in ENCODE regions. Genome Res 2007;17:746–59.
6. de Vooght KM, van Wijk R, van Solinge WW. GATA-1 binding sites in exon 1 direct erythroid-specific transcription of PPOX. Gene 2008;409: 83–91.
7. Zimmermann N, Colyer JL, Koch LE, Rothenberg ME. Analysis of the CCR3 promoter reveals a regulatory region in exon 1 that binds GATA-1. BMC Immunol 2005;6:7.
8. Kimura K, Wakamatsu A, Suzuki Y, Ota T, Nishikawa T, Yamashita R, et al. Diversification of transcriptional modulation: large-scale identification and characterization of putative alternative promoters of human genes. Genome Res 2006; 16:55–65.
9. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 2003;21:577–81.
10. Rockman MV, Wray GA. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol 2002;19:1991–2004.
11. Ladenvall P, Johansson L, Jansson JH, Jern S, Nilsson TK, Tjarnlund A, et al. Tissue-type plasminogen activator −7,351C/T enhancer polymorphism is associated with a first myocardial infarction. Thromb Haemost 2002;87:105–9.
12. Cooper DN. Human gene mutation in pathology and evolution. J Inherit Metab Dis 2002;25: 157–82.
13. Finishing the euchromatic sequence of the human genome. Nature (Lond) 2004;431:931–45.
14. Poncz M, Ballantine M, Solowiejczyk D, Barak I, Schwartz E, Surrey S. beta-Thalassemia in a Kurdish Jew: single base changes in the T-A-T-A box. J Biol Chem 1982;257:5994–6.
15. Hardison RC, Chui DHK, Giardine B, Riemer C, Patrinos GP, Anagnou N, et al. HbVar: a relational database of human hemoglobin variants and thalassemia mutations at the globin gene server. Hum Mutat 2002;19:225–33.
16. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, Grzelczak Z, et al. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science (Wash DC) 1989;245:1066–73.
17. Bienvenu T, Lacronique V, Raymondjean M, Cazeneuve C, Hubert D, Kaplan JC, Beldjord C. Three novel sequence variations in the 5′ upstream region of the cystic fibrosis transmembrane conductance regulator (CFTR) gene: two polymorphisms and one putative molecular defect. Hum Genet 1995;95:698–702.
18. McCarthy VA, Harris A. The CFTR gene and regulation of its expression. Pediatr Pulmonol 2005; 40:1–8.
19. Knight JC. Functional implications of genetic variation in non-coding DNA for disease susceptibility and gene regulation. Clin Sci (Lond) 2003;104: 493–501.
20. Wang DG, Fan JB, Siao CJ, Berno A, Young P, Sapolsky R, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science (Wash DC) 1998;280:1077–82.
21. Antonarakis SE, Krawczak M, Cooper DN. Disease-causing mutations in the human genome. Eur J Pediatr 2000;159:S173–8.

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022

22. Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, et al. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res 2000;28:316–9.
23. Schug J, Overton GC. TESS: Transcription element search software on the WWW. Pennsylvania: Computational Biology and Informatics Laboratory, School of Medicine, University of Pennsylvania; 1998. 10 p. Available from: http://www.cbil.upenn.edu/tess/techreports/1997/CBIL-TR-1997–1001-v0.0.pdf.
24. Fogel GB, Weekes DG, Varga G, Dow ER, Craven AM, Harlow HB, et al. A statistical analysis of the TRANSFAC database. Biosystems 2005; 81:137–54.
25. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 2003;20:1377–419.
26. Alam J, Cook JL. Reporter genes: application to the study of mammalian gene transcription. Anal Biochem 1990;188:245–54.
27. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. Scanning human gene deserts for long-range enhancers. Science (Wash DC) 2003;302:413.
28. Carey M, Smale ST. Transcriptional regulation in eukaryotes: concepts, strategies, and techniques. New York: Cold Spring Harbor Laboratory Press; 2000. 640 p.
29. Garner MM, Revzin A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic Acids Res 1981;9: 3047–60.
30. Orlando V, Strutt H, Paro R. Analysis of chromatin structure by in vivo formaldehyde cross-linking. Methods 1997;11:205–14.
31. Wray GA. The evolutionary significance of cis-regulatory mutations. Nat Rev Genet 2007;8: 206–16.
32. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet 2004;5:276–87.
33. Ludlow LB, Schick BP, Budarf ML, Driscoll DA, Zackai EH, Cohen A, Konkle BA. Identification of a mutation in a GATA binding site of the platelet glycoprotein Ibβ promoter resulting in the Bernard-Soulier syndrome. J Biol Chem 1996;271: 22076–80.
34. Manco L, Ribeiro ML, Máximo V, Almeida H, Costa A, Freitas O, et al. A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. Br J Haematol 2000;110: 993–7.
35. Van Wijk R, Van Solinge WW, Nerlov C, Beutler E, Gelbart T, Rijksen G, Nielsen FC. Disruption of a novel regulatory element in the erythroid-specific promoter of the human *PKLR* gene causes severe pyruvate kinase deficiency. Blood 2003;101: 1596–602.
36. Crossley M, Brownlee GG. Disruption of a C/EBP binding site in the factor IX promoter is associated with haemophilia B. Nature (Lond) 1990; 345:444–6.
37. Reijnen MJ, Sladek FM, Bertina RM, Reitsma PH. Disruption of a binding site for hepatocyte nuclear factor 4 results in hemophilia B Leyden. Proc Natl Acad Sci U S A 1992;89:6300–3.
38. Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res 1978;5: 3157–70.

Downloaded from https://academic.oup.com/clinchem/article/55/4/698/5629367 by guest on 20 August 2022